An annotated chromosome-scale reference genome for Eastern black-eared wheatear (*Oenanthe melanoleuca*)

Valentina Peona[1,*], Octavio Manuel Palacios-Gimenez[1,2,3*], Dave Lutgen[4,2,5,*], Remi André Olsen[6], Niloofar Alaei Kakhki[2], Pavlos Andriopoulos[7], Vasileios Bontzorlos[8], Manuel Schweizer[9,4], Alexander Suh[1,10], Reto Burri[5,4,3]

[1] Department of Organismal Biology – Systematic Biology, Science for Life Laboratory, Evolutionary Biology Centre, Uppsala University, 75236 Uppsala, Sweden

[2] Department of Population Ecology, Institute of Ecology and Evolution, Friedrich Schiller University Jena, 07743 Jena, Germany

[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany

[4] Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland

[5] Swiss Ornithological Institute, CH-6204 Sempach, Switzerland

[6] Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 17165 Solna, Sweden

[7] Section of Ecology and Systematics, Department of Biology, National and Kapodistrian University of Athens, 15772 Athens, Greece

[8] TYTO – Association for the Management and Conservation of Biodiversity in Agricultural Ecosystems, 41335 Larisa, Greece

[9] Natural History Museum Bern, 3005 Bern, Switzerland

[10] School of Biological Sciences, University of East Anglia, NR4 7TU Norwich, United Kingdom

* These authors contributed equally to the work.

Keywords
Birds, open-habitat chats, *Oenanthe melanoleuca*, *Oenanthe hispanica*-complex, transcriptome, repeat content, transposable elements

Running head
Eastern black-eared wheatear genome

Correspondence
Reto Burri, Swiss Ornithological Institute, Seerose 1, CH-6204 Sempach, Switzerland
Email: reto.burri@vogelwarte.ch

ORCiD
Valentina Peona: 0000-0001-5119-1837
Octavio Manuel Palacios-Gimenez: 0000-0002-1472-9949
Dave Lutgen: 0000-0003-0793-3930
Pavlos Andriopoulos: 0000-0002-5377-2974
Vasileios Bontzorlos: 0000-0002-1276-3385
Manuel Schweizer: 0000-0002-7555-8450

Alexander Suh: 0000-0002-8979-9992
Reto Burri: 0000-0002-1813-0079

Abstract

Pervasive convergent evolution and in part high incidences of hybridization distinguish wheatears (songbirds of the genus *Oenanthe*) as a versatile system to address questions at the forefront of research on the molecular bases of phenotypic and species diversification. To prepare the genomic resources for this venture, we here generated and annotated a chromosome-scale assembly of the Eastern black-eared wheatear (*O. melanoleuca*). This species is part of the *O. hispanica*-complex that is characterized by convergent evolution of plumage coloration and high rates of hybridization. The long-read-based male nuclear genome assembly comprises 1.04 Gb in 32 autosomes, the Z chromosome, and the mitogenome. The assembly is highly contiguous (contig N50: 12.6 Mb; scaffold N50: 70 Mb), with 96 % of the genome assembled at chromosome level and 95.5 % BUSCO completeness. The nuclear genome was annotated with 18,143 protein-coding genes and 31,333 mRNAs (annotation BUSCO completeness: 98.0 %), and about 10 % of the genome consists of repetitive DNA. The annotated chromosome-scale reference genome of Eastern black-eared wheatear provides a crucial resource for research into the genomics of adaptation and speciation in an intriguing group of passerines.

Introduction

Wheatears of the genus *Oenanthe* and their relatives – together referred to as "open-habitat chats" – are a group of songbirds that display several remarkable characteristics distinguishing them as a versatile system to address key questions on the evolution of phenotypes and formation of species. Many phenotypes, including multiple conspicuous colour ornaments, seasonal migration, and sexual dimorphism appear independently in multiple branches within open-habitat chats, suggesting a high incidence of convergent evolution (Alaei Kakhki et al. in press; Aliabadian et al. 2012; Schweizer et al. 2019). Furthermore, hybridization is observed in several species complexes and occurs at notably high rates in the *O. hispanica*-complex that consists of four currently recognized taxa (Schweizer et al. 2019): Western black-eared wheatear (*O. hispanica*), pied wheatear (*O. pleschanka*), cyprus wheatear (*O. cypriaca*), and Eastern black-eared wheatear (*O. melanoleuca*; **Fig. 1**). Pied and Eastern black-eared wheatear hybridize pervasively at the western shores of the Black Sea, in the Caucasus, and in the Alborz mountains of northern Iran (Haffer 1977; Panov 2005). The resulting introgression reaches beyond the hybrid zones (Schweizer et al. 2019), and hybrid zones themselves sport admixed phenotypes that display combinations of plumage colour phenotypes divergent between species (mantle and neck-side coloration) (Haffer 1977; Panov 2005). Finally, a phenotype divergently expressed between many wheatear species, black-or-white throat coloration, segregates as polymorphisms in three species of the *O. hispanica*-complex. Once a high-quality reference genome is available, this polymorphism and the recombination of mantle and neck-side coloration in hybrids provide an excellent opportunity to map these phenotypes to the genome (Buerkle and Lexer 2008) and study their convergent evolution across open-habitat chats. Furthermore, hybridization in

several geographic regions enables insights into common or idiosyncratic patterns of evolution under hybridization (Gompert et al. 2017).

Here, we describe the *de novo* assembly and annotation of a chromosome-scale reference genome for the Eastern black-eared wheatear (*O. melanoleuca*). The assembly includes models for 32 autosomes, the Z chromosome and the mitogenome that together cover 90 % of the k-mer-based genome size estimate (94 % with unplaced scaffolds included); it is highly contiguous with a scaffold N50 of 70 Mb and BUSCO completeness score of 95.5 %. This reference genome enables genomic research into the evolutionary history of phenotypic and species diversification in wheatears and their close relatives.

Material and Methods

Sampling, tissue preservation, and nucleic acid extraction

To obtain optimal starting material for a reference individual, we freshly sampled a male Eastern black-eared wheatear (*Oenanthe melanoleuca*) well outside known hybrid zones (Haffer 1977; Panov 2005) in Galaxidi, Greece (sampling permit no. 181968/989, issued by the Ministry of Environment and Energy, General Secretariat of Environment, General Directorate of Forests and Forest Environment, Directorate of Forest Management, Department of Wildlife and Game Management; export permit no. 55980/1575, Regional CITES management authority Attika). For this purpose, we sampled about 100 µl of blood from the brachial vein, and, after euthanizing the bird, we extracted all tissues possible. Tissues were immediately snap-frozen in liquid nitrogen. Throughout transportation and storage preceding DNA extraction, the samples were kept at a temperature below -80° C.

To obtain ultra-high molecular weight (UHMW) DNA from the reference individual, NGI Uppsala (Sweden) extracted DNA from the blood sample using the Bionano Prep™ Blood and Cell Culture DNA Isolation Kit (Bionano, San Diego, USA). Electrophoresis on a Femto Pulse instrument showed a mean DNA fragment length of about 200 kb, with fragments reaching up to 800 kb.

To prepare muscle tissue for Hi-C sequencing library preparation, we pulverized breast muscle tissue from the reference individual in a mortar. To avoid unfreezing of the tissue powder, the procedure was carried out in a climate chamber at 4°C under regular addition of liquid nitrogen.

To prepare RNA for full-length transcript sequencing, we extracted total RNA from eight snap-frozen tissues kept at -80°C (brain, breast muscle, heart, kidney, liver, lung, spleen, and testis) using the RNeasy Mini Kit (Qiagen; Hombrechtikon, Switzerland) according to the manufacturer's instructions. RNA quality was assessed with a Fragment Analyzer (Agilent). RNA from spleen showed considerable degradation and was excluded from further analyses.

De novo genome sequencing, and reference genome assembly and annotation

Assembly strategy and data acquisition

To obtain a chromosome-scale reference genome, our strategy largely followed the multiplatform approach recommended by Peona et al. (2021). In brief, it consisted of (i) a phased primary assembly based on long reads (ii) polishing and scaffolding of the primary assembly with linked-read

sequencing data, and (iii) scaffolding of the secondary assembly with proximity ligation (Hi-C) information.

To this end, we obtained a total of 215 Gb (unique coverage 151 Gb) Pacific Biosciences (PacBio) long-read sequence data, 54 Gb linked-read sequence data, and 83 Gb Hi-C data. NGI Uppsala (Sweden) prepared a PacBio library from UHMW DNA using the SMRTbell Template Prep Kit 1.0 and sequenced this library on 18 SMRT Cells 1M v3 on a PacBio Sequel instrument (Sequel Binding Kit 3.0, Sequel Sequencing Plate 3.0). PacBio long-read data was initially processed using SMRT Link v6. A linked-read sequencing library was prepared using the 10X Genomics Chromium Genomic Kit (from the same DNA extraction as used for PacBio sequencing; 10X Genomics, Inc., Pleasanton, CA, USA; Cat No. 120215), and a Hi-C library was prepared with the Dovetail Omni-C kit (Scotts Valley, CA, USA; Cat No. 21005). The linked-read and Hi-C libraries were prepared and sequenced on a NovaSeq 6000 instrument (S4 lane, 150 bp paired-end reads) at the facilities of NGI Stockholm (Sweden).

Genome size estimation

We estimated genome size by counting k-mer frequency of the quality-checked 10X Genomics linked reads. To this end, we first trimmed 22 bp from all 10X Genomics linked reads using fastp (Chen et al. 2018) to remove indices from R1 reads and keep symmetric read lengths for the R2 reads. We then counted k-mers of size 21 using jellyfish 2.2.10 (Marçais and Kingsford 2011) and used GenomeScope (Vurture et al. 2017) to estimate genome size from k-mer count histograms.

***De novo* genome assembly**

We assembled the PacBio long reads into the phased primary assembly using the Falcon Unzip 0.5 assembler (Chin et al. 2016), followed by polishing with Arrow 1.9.0. Before assembly polishing, we masked repeat regions of the phased primary assembly with RepeatMasker 4.1.0 (Smit et al. 1996-2010) using a custom repeat library (Boman et al. 2019; Peona et al. 2021; Suh et al. 2018; Weissensteiner et al. 2020) to make accurate assembly corrections without overcorrecting large repeats. We then polished the masked assembly with two rounds in Pilon v1.22 (Walker et al. 2014) with the parameter "--fix indels" using the reference individual's linked-read data. To purge duplicate scaffolds from the assembly, we ran purge_dups 1.2.5 (Guan et al. 2020) on the polished assembly. Prior to scaffolding with linked-read data, we split potential mis-assemblies with reference-individual linked-read data using Tigmint 1.2.4 (Jackman et al. 2018). With the aim to scaffold the polished remaining contigs, we applied ARCS 1.2.2 andLINKS 2.0.0 using the reference individual's linked-read data using default parameters (Warren et al. 2015; Yeo et al. 2018).

To further scaffold the assembly, we applied the 3D-DNA pipeline (Dudchenko et al. 2017) to join the sequences into chromosomes. We first used Juicer v.1.6 (Durand et al. 2016) to map Hi-C data against the contigs and to filter reads, and then ran the asm-pipeline v.180922 to generate a draft scaffolding.

Finally, we corrected mis-assemblies based on the visual inspection of the proximity map using Juicebox 2.13.06 (Robinson et al. 2018). The final chromosome-level assembly was polished with two additional rounds in Pilon as described above.

To assess homology of the assembled scaffolds with bird chromosomes, we aligned the final genome assembly to the genomes of collared flycatcher (*Ficedula albicollis*) (FicAlb1.5) (Kawakami

et al. 2014), zebra finch (taeGut3.2.4) (Warren et al. 2010), and chicken (GRCg6a) (Bellott et al. 2017) using D-Genies (Cabanettes and Klopp 2018). Chromosomes were named according to homology with these three genomes. In cases, such as chicken chromosomes 1 and 4 that are split to multiple chromosomes in songbirds, the nomenclature in the wheatear genome was adapted to the species whose homologous chromosome matched closest.

Mitogenome assembly

To assemble the mitochondrial genome, we used the MitoFinder 1.4 (Allio et al. 2020) and mitoVGP 2.2 (Formenti et al. 2021) pipelines with the published *Oenanthe isabellina* mitochondrial genome (Genbank Accession Number: NC_040290.1) as reference. We ran MitoFinder with the reference individual's short-read data (linked-read data but without making use of the linked-read haplotype information), and with mitoVGP we made joint use of the linked-read and long-read data. From MitoFinder we extracted the longest contig containing all 13 protein coding genes, two rRNA genes and 22 tRNAs annotated by MitoFinder as mitogenome assembly. We annotated both assemblies using the MITOS WebServer (http://mitos2.bioinf.uni-leipzig.de/index.py).

We then aligned both resulting assemblies with the mitogenomes of isabelline wheatear (*O. isabellina*, NC_040290.1) and northern wheatear (*O. Oenanthe*, MN356231.1) using MUSCLE (Edgar 2004) in MEGA X (Stecher et al. 2020) and generated a circular mitogenome map using CGView (Stothard and Wishart 2005).

Assembly quality evaluation

To evaluate assembly quality at each assembly step, we estimated basic assembly statistics using QUAST 5.0.2 (Gurevich et al. 2013) and evaluated the completeness of expected gene content in the assembly based on benchmarking universal single-copy orthologs (BUSCO) (Simão et al. 2015) with the avian dataset aves_odb10 (8,338 BUSCO) in BUSCO 5.0.0.

Repeat annotation

The final version of the genome assembly was used to *de novo* characterize both interspersed and tandem repeats. For interspersed repeats, we used RepeatModeler2 (Flynn et al. 2020) with the option "-LTR_struct" to obtain an improved characterisation of LTR retrotransposons which are commonly found in avian genomes (Boman et al. 2019; Kapusta and Suh 2017; Peona et al. 2021). The resulting library of raw consensus sequences was filtered from consensus sequences of tandem repeats (for which we ran a specific analysis; see below) and from protein-coding genes using the Snakemake pipeline repeatlib_filtering_workflow v0.1.0 (https://github.com/NBISweden/repeatlib_filtering_workflow).

For tandem repeats, we used RepeatExplorer2 (Novák et al. 2020) to search for satellite DNA (satDNA) sequences using the reference individual's 10X Genomics linked reads. Prior to RepeatExplorer2 graph-based clustering analysis, sequencing reads were pre-processed and checked by quality with FastQC (Babraham Bioinformatics: Cambridge 2012) using the public online platform at https://repeatexplorer.elixir-cerit-sc.cz. We processed the reads with the "quality trimming tool", "FASTQ interlacer on the paired end reads", "FASTQ to FASTQ converter", followed by

"RepeatExplorer2 clustering" with default parameters. Each reference sequence assembled by RepeatExplorer2 consisted of a monomer of the satDNA consensus sequence. The relative genomic abundance and nucleotide divergence (Kimura-2-parameter distance) of each detected satDNA were estimated by sampling four million read pairs and aligning them to the satDNA library with RepeatMasker 4.1.0 (Smit et al. 1996-2010). The sampled reads were mapped to dimers of satDNA consensus sequences, and for smaller satDNAs, several monomers were concatenated until reaching roughly 150 bp array length. The resulting RepeatMasker *.align* file was then parsed to the script *calcDivergenceFromAlign.pl* from RepeatMasker utils. The relative abundance of each satDNA sequence was then estimated as the proportion of nucleotides aligned with the reference sequence with respect to the total Illumina library size.

The RepeatModeler2 library was then merged with the satDNA library produced here and with known avian consensus sequences of transposable elements from Repbase (Bao et al. 2015), Dfam (Storer et al. 2021, 2021), flycatcher, blue-capped cordon-bleu, hooded crow, and paradise crow (Boman et al. 2019; Peona et al. 2021; Suh et al. 2018; Weissensteiner et al. 2020). This library was then used to annotate the genome assembly with RepeatMasker (Smit et al. 1996-2010). The annotation produced was processed with the script *calcDivergenceFromAlign.pl* from RepeatMasker utils to calculate the divergence between repeats and their consensus sequences using the Kimura 2-parameter distance corrected for the presence of CpG sites.

Full-length transcript sequencing and genome annotation

We aimed to establish a high-quality genome annotation based on full-length transcripts. To this end, for each of the abovementioned seven tissues, the NGS platform of the University of Berne, Switzerland, prepared an Iso-Seq library using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences). These seven libraries were then sequenced on three separate SMRT cells 8M, sequencing twice five tissues (brain and testis, lung, muscle, and heart) and once two tissues (liver and kidney) per SMRT cell. Sequencing of these SMRT cells was conducted on a Pacific Biosciences Sequel II instrument at the Genomic Technologies Facility in Lausanne, Switzerland. As the libraries underloaded, five libraries (all but liver and kidney) were jointly sequenced on an additional SMRT cell 8M on a Pacific Biosciences Sequel IIe at the NGS platform of the University of Berne.

Circular consensus sequences (CCS), full-length non-chimeric transcripts, and polished high- and low-quality transcripts were obtained by the NGS platform at the University of Bern separately for each run using the Isoseq 3 pipeline (ICS v10.1). Polished full-length isoforms for each sequencing run were merged by tissue and then separately mapped to the reference genome using Minimap v2.2 (-ax splice) (Li 2018, 2021). Transcriptome annotations were generated by first collapsing redundant transcripts using TAMA collapse (-x no_cap), before generating open reading frame (ORF) and nonsense-mediated mRNA decay (NMD) predictions using the scripts implemented in TAMA-GO (Kuo et al. 2020) for each of the seven tissues. We then evaluated tissue-specific transcriptome completeness using BUSCO (Simão et al. 2015) with the avian dataset aves_odb10 (8'338 BUSCO) in BUSCO 5.0.0. Additional transcriptome annotation statistics were obtained using the *agat_sp_statistics.pl* script implemented in the AGAT toolkit (Dainat 2019).

We annotated the repeat soft-masked genome using GeMoMa 1.9 (Keilwagen et al. 2018; Keilwagen et al. 2019), a homology-based gene prediction tool. This tool is based on the annotation

of protein-coding genes and intron position conservation in a reference genome to predict the annotation of protein-coding genes in the target genome. We used the genomes of chicken (GCA_016699485.1; International Chicken Genome Sequencing Consortium 2004), zebra finch (GCA_003957565.2; Warren et al. 2010), silvereye (GCA_001281735.1; Cornetti et al. 2015), and collared flycatcher (GCA_000247815.2; Ellegren et al. 2012; Kawakami et al. 2014) as references for the homology-based gene prediction, along with the reference individual's transcriptome obtained from Iso-Seq data to incorporate RNA evidence for the splice prediction. Using the Extract RNA-seq Evidence tool implemented in GeMoMa, we obtained intron position and coverage. This information was fed into the GeMoMa pipeline (GeMoMa.m=200000, AnnotationFinalizer.r=SIMPLE, pc=true, and o=true) to obtain predicted protein-coding gene models. To account for redundancies/duplicates resulting from the predicted protein-coding genes potentially stemming from each of the four reference species, genome annotation completeness was assessed by recomputing BUSCO using the BUSCOrecomputer tool in GeMoMa.

Functional annotation of protein-coding genes was obtained with InterProScan 5.59 (Jones et al. 2014; Paysan-Lafosse et al. 2022). InterProScan ran with the following settings: *-goterms -iprlookup -appl CDD, COILS, Gene3D, HAMAP, MobiDBLite, PANTHER, Pfam, PIRSF, PRINTS, PROSITEPATTERNS, PROSITEPROFILES, SFLD, SMART, SUPERFAMILY, TIGRFAM*). Predicted protein-coding genes were further annotated through a protein Blast search (-evalue 0.000001, -seg yes, -soft_masking true, -lcase_masking) against the Swiss-Prot database (Uniprot Consortium 2019). We then merged the predicted protein-coding gene models and the functional annotation using the *agat_sp_manage_functional_annotation.pl* script, obtained summary statistics using *agat_sp_statistics.pl* and *agat_sp_functional_statistics.pl*, both implemented in the AGAT toolkit. Gene ontology (GO-terms) were visualised with WEGO 2.0 (wego.genomics.cn).

Results and Discussion

Nuclear genome assembly

The polished, unzipped primary assembly contained a total of 1,681 contigs, of which all were >25 kb long and 1,610 were >50 kb long (**Tab. 1**). Total assembly length was 1.29 Gb, with the longest contig spanning 45.3 Mb, contig N50 of 8.6 Mb, and half of the assembly placed in 35 contigs. Avian BUSCO were 96.9 % complete, with 90.6 % being single-copy genes (**Tab. 1**).

Purging duplicated contigs resulted in an assembly constituted of 381 contigs with a total assembly length of 1.04 Gb, contig N50 of 13.5 Mb and half of the assembly placed in 23 contigs (**Tab. 1**). After this step, BUSCO completeness remained at 96.4 %, but an improvement to nearly 96 % single-copy BUSCOs was achieved (**Tab. 1**).

Starting from an already highly contiguous assembly, the linked-read data did not yield any scaffolding improvement. Still, Tigmint detected several supposed mis-assemblies and split the assembly into 451 scaffolds. However, an alignment of the original contigs in D-Genies (Cabanettes and Klopp 2018) showed that all but one of the original contigs (see below) were collinear with the collared flycatcher genome. Given this result and that the proximity ligation data would correct mis-assemblies in subsequent steps, we decided to keep the original contigs except for one aligning to flycatcher chromosomes 2 and 3. For the latter contig, we used the output of Tigmint that split the

contig in line with the alignment. The two split parts covered all but 12,527 bp of the original contig. Visual inspection of the missing sequence showed that it almost entirely consisted of repeats. We left this sequence in the assembly as a separate contig.

The proximity ligation information obtained through Hi-C scaffolding corrected a number of scaffolds, resulting in a higher number of scaffolds (588) than the number of contigs it started from (383). However, the scaffolding yielded a highly contiguous chromosome-scale assembly (N50, 69.6 Mb; L50, 6) with BUSCO completeness of still >96 % and almost all BUSCOs in single copy (**Tab. 1**). This final assembly contained all macrochromosomes and the majority of microchromosomes usually found in the latest generation of avian genome assemblies (Kapusta et al. 2017; Peona 2021; Rhie et al. 2021). 96 % of the assembly was placed into chromosome models, and the chromosome-only assembly covered still 95.5 % of BUSCO (**Tab. 1**).

The final assembly length closely matched the one of previous linked-read-based assemblies of the same species and closely related ones (Lutgen et al. 2020; Schweizer et al. 2019). The genome size estimated from the k-mer distribution of linked reads sequence was between 1.105 and 1.106 Gb, with 0.925-0.926 Gb of unique and 0.179-0.180 Gb (16 %) repeat sequence and 0.75-0.76 % heterozygosity (GenomeScope model fit 98-99 %). The full final reference genome assembly thus covered 94 % of the genome size estimate, with 90 % of the estimated genome size placed in chromosomes. 96 % of the assembly were placed in 33 chromosomes with homologs in collared flycatcher, zebra finch and chicken, according to which we adapted the chromosome nomenclature. The differences in genome size estimates based on the k-mer approach and the genome assembly length is likely the result of highly repetitive sequences (e.g., centromeres, telomeres, satDNAs) that collapsed during the assembly process (Peona et al. 2018). Assembly contiguity and completeness (as judged by BUSCO scores) of the *O. melanoleuca* assembly compared favourably to other songbird genome assemblies (**Tab. 2**).

Mitogenome assembly

MitoFinder and MitoVGP assembled mitogenomes of 16,944 bp and 18,631 bp length, respectively. The mitochondrial contigs assembled by the two pipelines were congruent, except for 9 single base pair mismatches, for a 1,827 bp long insert in the MitoVGP assembly and of a 141 bp long insert in the MitoFinder assembly. We decided to not consider either of these inserts in the final mitogenome assembly for the following reasons. First, neither of the inserts was observed in the mitogenomes of isabelline and northern wheatear. For the long insert in the MitoVGP assembly, moreover, the coverage of short reads mapped to the MitoVGP assembly was strongly reduced (**Fig. S1**), and the insertion constituted a partial duplication of *nd6*, duplications of two tRNAs (Glu, Pro) and a partial duplication of the control region likely caused by an assembly artefact. The short insert in the MitoFinder assembly was not observed in the other wheatear mitogenomes, and if real, we would expect long reads to cover this insert. Because base calling based on short reads is expected to have higher quality, we retained the MitoFinder assembly, but without the 141 bp insert as final mitogenome.

The final mitogenome (as also both original assemblies) contained all 13 protein-coding genes, two rRNAs, and 22 tRNAs (**Fig. 2**). All genes, except eight tRNAs and *nd6*, were located on the

heavy DNA strand. Both gene order and strandedness were concordant with those observed in northern wheatear (*O. oenanthe*) (Wang et al. 2020).

Repetitive element annotation

The *de novo* identification of repetitive elements resulted in the characterisation of 572 raw consensus sequences from RepeatModeler2 and 16 satellite DNA consensus sequences from RepeatExplorer2. The consensus sequences from RepeatModeler2 were filtered from tandem repeats and protein-coding genes. This resulted in a final library of 477 consensus sequences (**File S1**). Among these consensus sequences, RepeatModeler2 classified 226 sequences as LTR retrotransposons, 98 as LINE retrotransposons, 21 as DNA transposons, 5 as SINE retrotransposons, and 112 sequences were unclassified ("unknown").

The genome assembly annotation run with RepeatMasker using the repeat library produced here and merged with already known avian repeats showed that ~10 % of the assembled genome is repetitive (**Fig. 3A**, **Tab. S1, File S2**). This finding indicates that many repeats collapsed during the genome assembly process. An example of this were satDNAs that represented ~0.8 % of the sequenced reads but only < 0.3 % of the genome assembly, suggesting that satDNA repeats (such as in (peri-)centromeric and (sub-)telomeric regions) are the most collapsed repeats. Most of the repeats annotated were LTR and LINE retrotransposons (**Fig. 3A**). While it is common to find LINEs as most abundant TEs in avian genomes (Galbraith et al. 2021; Kapusta and Suh 2017; Manthey et al. 2018; Peona, Blom et al. 2021), it is less common to find so similar percentages of LINE and LTR retrotransposons. This is especially true for a male genome assembly such as the present one here that does not include the W chromosome which is highly enriched in LTRs and acts as a refugium for most of the full-length genomic LTR elements in birds (Peona et al. 2021; Warmuth et al. 2022). The transposable element landscape (**Fig. 3B**) suggests that LINE retrotransposons experienced a drop in their genomic accumulation in recent times (0-5 % divergence; **Fig. 3B**), whereas LTR retrotransposons kept accumulating at the same rate. Such a recent replacement of LINE retrotransposon activity with a diversity of LTR retrotransposons has been noted in other songbirds and seems to have occurred independently in the so far analysed passerine families, i.e., estrildid finches (Warren et al. 2010, Boman et al. 2019), flycatchers (Suh et al. 2018), crows (Weissensteiner et al. 2020), and birds-of-paradise (Peona et al. 2021). Finally, the satDNA landscape (**Fig. 3B**) shows that satDNA arrays experienced differential amplification in copies number in recent times (0-10 % divergence), implying fast evolution of this genomic fraction in the genome (Peona et al. 2022).

Transcriptome sequencing, genome annotation, and gene function prediction

Iso-Seq sequencing yielded a total of 4,627,382 CCS reads (125,633-1,087,892 reads per tissue, **Tab. 3**). This resulted in numbers of high-quality isoforms ranging from 16,078 to 80,600 per tissue. On average 8'833 genes were predicted per tissue, ranging from 4,772 in muscle to 10,924 in liver. Transcriptome completeness evaluated through BUSCO ranged from 31.2 % to 57.5 % complete BUSCO per tissue (**Tab. 3**).

The Iso-Seq transcriptomes were then used as splice evidence in GeMoMa to perform a predominantly homology-based annotation of the reference genome. We predicted 18,143 protein-coding genes with a total of 320,754 exons and 289,421 introns. The number of exons, CDS, and introns was higher for our *O. melanoleuca* annotation compared to the annotations of other songbirds, such as *Junco hyemalis*, *Fringilla coelebs*, *Melospiza melodia*, *Taeniopygia guttata*, *Ficedula albicollis*, *Manacus vitellinus,* and *Geospiza fortis* (**Tab. 2**). Mean gene length, CDS length, exon length, and number of exons per gene, on the other hand, were in the range of values obtained for the abovementioned songbird annotations (**Tab. 3**). 17,'553 (96.7 %) of the 18,'143 predicted genes were annotated with protein families or function assignment. 12,'472 (68.7 %) genes obtained a GO term assignment through InterProScan. The most abundant GO terms were associated with "cell part", "cell" and "membrane" in the cellular component category, "binding" in the molecular function category and "cellular metabolic process" or "metabolic process" in the biological process category (**Fig. S2**). BUSCO completeness of the final annotation as judged from avian BUSCO (n=8,338) was 98.0 %, with 97.4 % single copy BUSCO, 0.6 % duplicated BUSCO, 0.6 % fragmented BUSCO, and 1.5 % missing BUSCO. This suggests an accurate and rather complete annotation.

Data Availability

All data, including the assembly, its annotation, and the original sequencing data are available on the European Nucleotide Archive under project assession PRJNA937434. Code for the repeat analysis is available on https://github.com/ValentinaBoP/WheatearGenomeAnalysis.

Supplemental Material is available at figshare: https://doi.org/10.25387/g3.22209697.

Acknowledgements

We warmly thank Marta Burri for preparing RNA, and Giulio Formenti, Remi Allio, and Lauren Coombe for support with computational questions. We are indebted to NGI Uppsala, namely Mai-Britt Mosbech and Olga Vinnere Pettersson for UHMW DNA extraction and long-read sequencing and Ignas Bunikis for running the primary assembly, as well as NGI Stockholm for the preparation of linked-read and Hi-C sequencing data. Finally, we thank the NGS platform at the University of Berne, namely Pamela Nicholson and Catia Coito for the preparation of Iso-Seq data. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) and the High-Performance Computing Cluster EVE, a joint effort of the Helmholtz Centre for Environmental Research (UFZ) and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-

Leipzig. We thank the administration and support staff of EVE: Thomas Schnicke and Ben Langenberg (UFZ), and Christian Krause (iDiv).

Conflict of Interest

The authors declare no conflict of interest.

Funder Information

The present project was supported by German Research Foundation (DFG), grant number BU3456/3-1 to RB and the National Research Fund (FNR) Luxembourg, grant number 14575729, to DL; OMPG was supported by the Swedish Research Council Vetenskapsrådet (grant number 2020-03866); VP was supported via grants to AS from the Swedish Research Council Vetenskapsrådet (grant number 2020-04436) and the Swedish Research Council Formas (2017-01597). NAK was supported by a Georg Foster Research Stipend of the Alexander von Humboldt Foundation. Part of the analyses were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Literature cited

Alaei Kakhki N, Schweizer M, Lutgen D, Bowie RCK, Shirihai H, Suh A, Schielzeth H, Burri R. In press. A phylogenomic assessment of processes underpinning convergent evolution in open-habitat chats. Mol. Biol. Evol (in press, bioRxiv doi:10.1101/2022.06.21.496980).

Aliabadian M, Kaboli M, Förschler MI, Nijman V, Chamani A, Tillier A, Prodon R, Pasquet E, Ericson PGP, Zuccon D. 2012. Convergent evolution of morphological and ecological traits in the open-habitat chat complex (Aves, Muscicapidae: Saxicolinae). Molecular Phylogenetics and Evolution. 65(2012):35–45.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular ecology resources. 20(4):892–905. eng. doi:10.1111/1755-0998.13160.

Babraham Bioinformatics: Cambridge. 2012. FastQC; Version 0.10.1: A Quality Control Tool for High throughput Sequence Data; Babraham Bioinformatics. Cambridge, UK.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 6:11. eng. doi:10.1186/s13100-015-0041-9.

Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, Galkina S, Pyntikova T, Koutseva N, Graves T, et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. Nature genetics. 49(3):387–394. eng. doi:10.1038/ng.3778.

Boman J, Frankl-Vilches C, da Silva dos Santos, Michelly, de Oliveira, Edivaldo H. C., Gahr M, Suh A. 2019. The Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch. Genes. 10(4):301.

Boman J, Frankl-Vilches C, da Silva dos Santos, Michelly, Oliveira EHC de, Gahr M, Suh A, da Silva dos Santos, Michelly. 2019. The Genome of Blue-Capped Cordon-Bleu Uncovers Hidden Diversity of LTR Retrotransposons in Zebra Finch. Genes. 10(4). eng. doi:10.3390/genes10040301.

Buerkle CA, Lexer C. 2008. Admixture as the basis for genetic mapping. Trends in Ecology & Evolution. 23(12):686–694. doi:10.1016/j.tree.2008.07.008.

Cabanettes F, Klopp C. 2018. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ. 6:e4958. eng. doi:10.7717/peerj.4958.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17):i884-i890. eng. doi:10.1093/bioinformatics/bty560.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Meth. 13(12):1050–1054. doi:10.1038/nmeth.4035.

Cornetti L, Valente LM, Dunning LT, Quan X, Black RA, Hébert O, Savolainen V. 2015. The Genome of the "Great Speciator" Provides Insights into Bird Diversification. Genome Biol Evol. 7(9):2680–2691. eng. doi:10.1093/gbe/evv168.

Dainat J. 2019. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. (Version v0.7.0). Zenodo. doi:10.5281/zenodo.3552717.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science (New York, N.Y.). 356(6333):92–95. eng. doi:10.1126/science.aal3327.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell systems. 3(1):95–98. eng. doi:10.1016/j.cels.2016.07.002.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32(5):1792–1797. doi:10.1093/nar/gkh340.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature. 491(7426):756–760. doi:10.1038/nature11584.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the United States of America. 117(17):9451–9457. eng. doi:10.1073/pnas.1921046117.

Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, Brown S, Capodiferro MR, Al-Ajli FO, Ambrosini R, et al. 2021. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 22(1):120. eng. doi:10.1186/s13059-021-02336-9.

Friis G, Vizueta J, Ketterson ED, Milá B. 2022. A high-quality genome assembly and annotation of the dark-eyed junco Junco hyemalis, a recently diversified songbird. G3 (Bethesda). 12(6). eng. doi:10.1093/g3journal/jkac083.

Galbraith JD, Kortschak RD, Suh A, Adelson DL. 2021. Genome Stability Is in the Eye of the Beholder: CR1 Retrotransposon Activity Varies Significantly across Avian Diversity. Genome Biol Evol. 13(12). eng. doi:10.1093/gbe/evab259.

Gompert Z, Mandeville EG, Buerkle CA. 2017. Analysis of Population Genomic Data from Hybrid Zones. Annual Review of Ecology, Evolution, and Systematics; [accessed 2017 Aug 21]. 48:207–229. doi:10.1146/annurev-ecolsys-110316-022652.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 36(9):2896–2898. eng. doi:10.1093/bioinformatics/btaa025.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 29(8):1072–1075. eng. doi:10.1093/bioinformatics/btt086.

Haffer J. 1977. Secondary contact zones of birds in Northern Iran. Bonner Zoologische Monographien. (10):1–64.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432(7018):695–716. eng. doi:10.1038/nature03154.

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJM, et al. 2018. Tigmint: Correcting assembly errors using linked reads from large molecules. BMC bioinformatics. 19(1):393. eng. doi:10.1038/nmeth.4035.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 30(9):1236–1240. eng. doi:10.1093/bioinformatics/btu031.

Kapusta A, Suh A. 2017. Evolution of bird genomes—a transposon's-eye view. Annals of the New York Academy of Sciences. 1389(1):164–185. doi:10.1111/nyas.13295.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences of the United States of America. 114(8):E1460-E1469.

Kawakami T, Backström N, Burri R, Husby A, Olason P, Rice AM, Ålund M, Qvarnström A, Ellegren H. 2014. Estimation of linkage disequilibrium and interspecific gene flow in Ficedula flycatchers by a newly developed 50k single-nucleotide polymorphism array. Molecular Ecology Resources. 14(6):1248–1260. doi:10.1111/1755-0998.12270.

Kawakami T, Smeds L, Backström N, Husby A, Qvarnström A, Mugal CF, Olason P, Ellegren H. 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Molecular Ecology. 23(16):4035–4058. doi:10.1111/mec.12810.

Keilwagen J, Hartung F, Grau J. 2019. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods Mol Biol. 1962:161–177. eng. doi:10.1007/978-1-4939-9173-0_9.

Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC bioinformatics. 19(1):189. eng. doi:10.1186/s12859-018-2203-5.

Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. 2020. Illuminating the dark side of the human transcriptome with long read transcript sequencing. BMC genomics. 21(1):751. eng. doi:10.1186/s12864-020-07123-7.

Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 34(18):3094–3100. eng. doi:10.1093/bioinformatics/bty191.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 37(23):4572–4574. eng. doi:10.1093/bioinformatics/btab705.

Lutgen D, Ritter R, Olsen R-A, Schielzeth H, Gruselius J, Ewels P, García JT, Shirihai H, Schweizer M, Suh A, et al. 2020. Linked-read sequencing enables haplotype-resolved resequencing at population scale. Molecular Ecology Resources. 20(5):1311–1322. doi:10.1111/1755-0998.13192.

Manthey JD, Moyle RG, Boissinot S. 2018. Multiple and Independent Phases of Transposable Element Amplification in the Genomes of Piciformes (Woodpeckers and Allies). Genome Biol Evol. 10(6):1445–1456. eng. doi:10.1093/gbe/evy105.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27(6):764–770. eng. doi:10.1093/bioinformatics/btr011.

Novák P, Neumann P, Macas J. 2020. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nature Protocols. 15(11):3745–3776. eng. doi:10.1038/s41596-020-0400-y.

Panov EN. 2005. Wheaters of the Palearctic. Ecology, Behaviour and Evolution of the genus *Oenanthe*. Sofia-Moscow: Pensoft (WIlson MG, editor. Series Faunistica).

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, et al. 2022. InterPro in 2022. Nucleic acids research. eng. doi:10.1093/nar/gkac993.

Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, Liachko I, Haryoko T, Jønsson KA, Zhou Q, et al. 2021. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. Molecular ecology resources. 21(1):263–286. eng. doi:10.1111/1755-0998.13252.

Peona V, Kutschera VE, Blom MPK, Irestedt M, Suh A. 2022. Satellite DNA evolution in Corvoidea inferred from short and long reads. Molecular Ecology. eng. doi:10.1111/mec.16484.

Peona V, Palacios-Gimenez OM, Blommaert J, Liu J, Haryoko T, Jønsson KA, Irestedt M, Zhou Q, Jern P, Suh A. 2021. The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 376(1833):20200186. eng. doi:10.1098/rstb.2020.0186.

Peona V, Weissensteiner MH, Suh A. 2018. How complete are "complete" genome assemblies?: —An avian perspective. Molecular Ecology Resources. 18(6):1188–1195. doi:10.1111/1755-0998.12933.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 592(7856):737–746. eng. doi:10.1038/s41586-021-03451-0.

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell systems. 6(2):256-258.e1. eng. doi:10.1016/j.cels.2018.01.001.

Schweizer M, Warmuth V, Alaei Kakhki N, Aliabadian M, Förschler MI, Shirihai H, Suh A, Burri R. 2019. Parallel plumage color evolution and pervasive hybridization in wheatears. Journal of Evolutionary Biology. 32(1):100–110.

Schweizer M, Warmuth VM, Alaei Kakhki N, Aliabadian M, Förschler M, Shirihai H, Ewels P, Gruselius J, Olsen R-A, Schielzeth H, et al. 2019. Genome-wide evidence supports mitochondrial relationships and pervasive parallel phenotypic evolution in open-habitat chats. Molecular Phylogenetics and Evolution. 139:106568.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics; [accessed 12/25/2019]. 31(19):3210–3212. doi:10.1093/bioinformatics/btv351.

Smit AFA, Hubley R, Green P. 1996-2010. RepeatMasker Open 3.3.0.

Stecher G, Tamura K, Kumar S. 2020. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Mol Biol Evol. 37(4):1237–1239. eng. doi:10.1093/molbev/msz312.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA. 12(1):2. eng. doi:10.1186/s13100-020-00230-y.

Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. Bioinformatics. 21(4):537–539. eng. doi:10.1093/bioinformatics/bti054.

Suh A, Smeds L, Ellegren H. 2018. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. Molecular Ecology. in press. doi:10.1111/mec.14439.

Uniprot Consortium. 2019. UniProt: A worldwide hub of protein knowledge. Nucleic acids research. 47(D1):D506-D515. eng. doi:10.1093/nar/gky1049.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics. 33(14):2202–2204. eng. doi:10.1093/bioinformatics/btx153.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 9(11):e112963. eng. doi:10.1371/journal.pone.0112963.

Wang E, Zhang D, Braun MS, Hotz-Wagenblatt A, Pärt T, Arlt D, Schmaljohann H, Bairlein F, Lei F, Wink M. 2020. Can Mitogenomes of the Northern Wheatear (Oenanthe oenanthe) Reconstruct Its Phylogeography and Reveal the Origin of Migrant Birds? Sci Rep. 10(1):9290. eng. doi:10.1038/s41598-020-66287-0.

Warmuth VM, Weissensteiner MH, Wolf JBW. 2022. Accumulation and ineffective silencing of transposable elements on an avian W Chromosome. Genome research. 32(4):671–681. eng. doi:10.1101/gr.275465.121.

Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, Birol I. 2015. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. GigaScience. 4:35. eng. doi:10.1186/s13742-015-0076-3.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunster A, Searle S, White S, Vilella AJ, Fairly S, et al. 2010. The genome of a songbird. Nature. 464(7289):757–762.

Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al. 2020. Discovery and population genomics of structural variation in a songbird genus. Nat Commun. 11(1):3403. eng. doi:10.1038/s41467-020-17195-4.

Yeo S, Coombe L, Warren RL, Chu J, Birol I. 2018. ARCS: Scaffolding genome drafts with linked reads. Bioinformatics. 34(5):725–731. eng. doi:10.1093/bioinformatics/btx675.

Table 1. Assembly statistics for different versions of the *O. melanoleuca* genome.

		Falcon unzip, Arrow	+ Pilon, purge_dups	+ Tigmint	+ 3D DNA (all)	+ 3D DNA (chrom)
Basic stats	No. contigs/scaffolds*	1,681	381	383	588*	32*
	No. contigs/scaffolds* > 50 kb	1,610	347	348	143*	31*
	Assembly length (Gb)	1.29	1.04	1.04	1.04*	1.00*
	Contig/scaffold* N50 (Mb)	8.6	13.5	12.6	69.6*	69.7*
	Contig/scaffold* L50	35	23	24	6*	5*
	Largest contig/scaffold* (Mb)	45.3	45.3	45.3	148.4*	148.4*
BUSCO	Complete (%)	96.9	96.4	96.4	96.2	95.5
	Complete single-copy (%)	90.6	95.9	95.9	95.7	95.1
	Complete duplicated (%)	6.3	0.5	0.5	0.5	0.4
	Fragmented (%)	0.7	0.7	0.7	0.9	0.9
	Missing (%)	2.4	2.9	2.9	2.9	3.6

* Where numbers concern scaffolds instead of contigs, this is indicated by an asterisk.

Table 2. Comparison of genome assembly and annotation summary statistics of *Oenanthe melanoleuca* with other songbird species (*Junco hyemalis*, *Fringilla coelebs*, *Melospiza melodia*, *Taeniopygia guttata*, *Ficedula albicollis*, *Manacus vitellinus*, and *Geospiza fortis*). Modified from Friis et al. (2022).

	Oenanthe	***Junco***	***Fringilla***	***Melospiza***	***Taeniopygia***	***Ficedula***	***Manacus***	***Geospiza***
Genome assembly length (Gb)	1.04	0.99	0.99	1.36	1.22	1.1	1.17	1.04
Genome contig N50 (kb)	7,700	75	67	8,300	38	410	194	30
Genome BUSCO scores (%) C	95.5	95.4	94.1	87.9	93.8	96.5	96.1	96.0
S	95.1	95.2	93.8	87.3	91.9	96	94.6	95.6
D	0.4	0.2	0.3	0.6	1.9	0.5	1.5	0.4
F	0.9	1.6	2.0	7.2	2.3	0.8	1	1.2
M	3.6	3.0	4.0	5.0	3.9	2.7	2.9	2.8
No. of genes	18,143	19,026	17,703	15,086	17,561	16,763	18,976	14,399
Mean gene length (bp)	28,23218	15,402	15,818	14,457	26,458	31,394	27,847	30,164
No. of CDS	31,333	23,245	17,703	15,086	17,561	16,763	18,976	14,399
Mean CDs length (bp)	1682	1,647	1,679	1,325	1,677	1,942	1,929	1,766
No. of exons	320,754	229,210	221,872	131,940	171,767	189,043	190,390	164,721
Mean exon length (bp)	164	167	165	153	255	253	264	195
Mean no. exons/gene	102	9.9	10.2	8.7	10.3	12.2	11.5	11.4
No. of introns	289,421	205,965	200,041	116,724	153,909	171,236	171,089	149,563

BUSCO parameters are C: complete genes; S: complete and single-copy genes; D: complete and duplicated genes; F: fragmented genes; M: missing genes.

1

Table 3. Iso-Seq data characterization and transcriptome completeness.

		Brain	Heart	Kidney	Liver	Lung	Muscle	Testis
Transcriptome	No. of CCS reads	847,617	253,468	723,158	1,087,892	1,061,936	125,633	527,678
	High-quality isoforms	73,422	80,600	45,097	47,491	28,508	16,078	44,605
	Low-quality isoforms	734	844	616	384	151	94	284
	No. of genes	10,449	10,448	9,063	10,924	6,564	4,772	9,613
	Mean gene length (bp)	24,193	20,119	16,350	15,125	18,528	17,397	17,415
	No. of CDS	27,449	28,747	25,790	27,202	13,551	8,447	23,009
	Mean CDS length (bp)	972	985	932	823	894	980	960
	No. of exons	231,169	222,791	235,989	194,325	108,084	69,859	184,794
	Mean exon length (bp)	246	248	223	225	221	224	209
	Mean no. of exons/mRNA	8.4	8.2	7.9	7.1	8.0	8.3	8.0
BUSCO	Complete (%)	56.80	57.50	48.30	49.40	38.30	31.20	49.3
	Single-copy (%)	40.30	39.50	33.60	34.70	31.1	27.00	34.6
	Duplicated (%)	16.50	18.00	14.70	14.70	7.20	4.20	14.70
	Fragmented (%)	2.90	2.10	2.60	3.20	2.00	1.10	2.30
	Missing (%)	40.30	40.60	49.10	47.40	59.70	67.70	48.40

Figure Legends

Figure 1. Eastern black-eared wheatear (*Oenanthe melanoleuca*). The species sports a white-throated (left; Agii Pantes, Greece, June 2022) and a black-throated phenotype (right; Lesvos, Greece, May 2017) in males. © Reto Burri

Figure 2. Circular sketch map of the *O. melanoleuca* mitogenome assembly. The outer circle shows coding sequences (purple), rRNAs (pink), and tRNAs (red). The black trace on the middle circle indicates GC content. On the inner circle, positive and negative GC skews in nucleotide composition are indicated by green and magenta, respectively.

Figure 3. Repeat annotation landscapes. **A**) Pie-chart summarizing the transposable element content annotated in the genome assembly. **B**) Transposable element landscape. The divergence between interspersed repeat copies and their consensus sequences is shown on the X-axis as genetic distance calculated using the Kimura 2-parameter distance. The percentage of the genome assembly occupied by transposable elements is shown on the Y-axis. **C**) Satellite DNA landscape. The divergence between the satellite DNA consensus sequences and sequences annotated in the short-read library is shown on the X-axis as genetic distance calculated using the Kimura 2-parameter distance. The percentage of the genome (short reads) annotated as satellite DNA is shown on the Y-axis.

Figure 1
160x53 mm (x DPI)

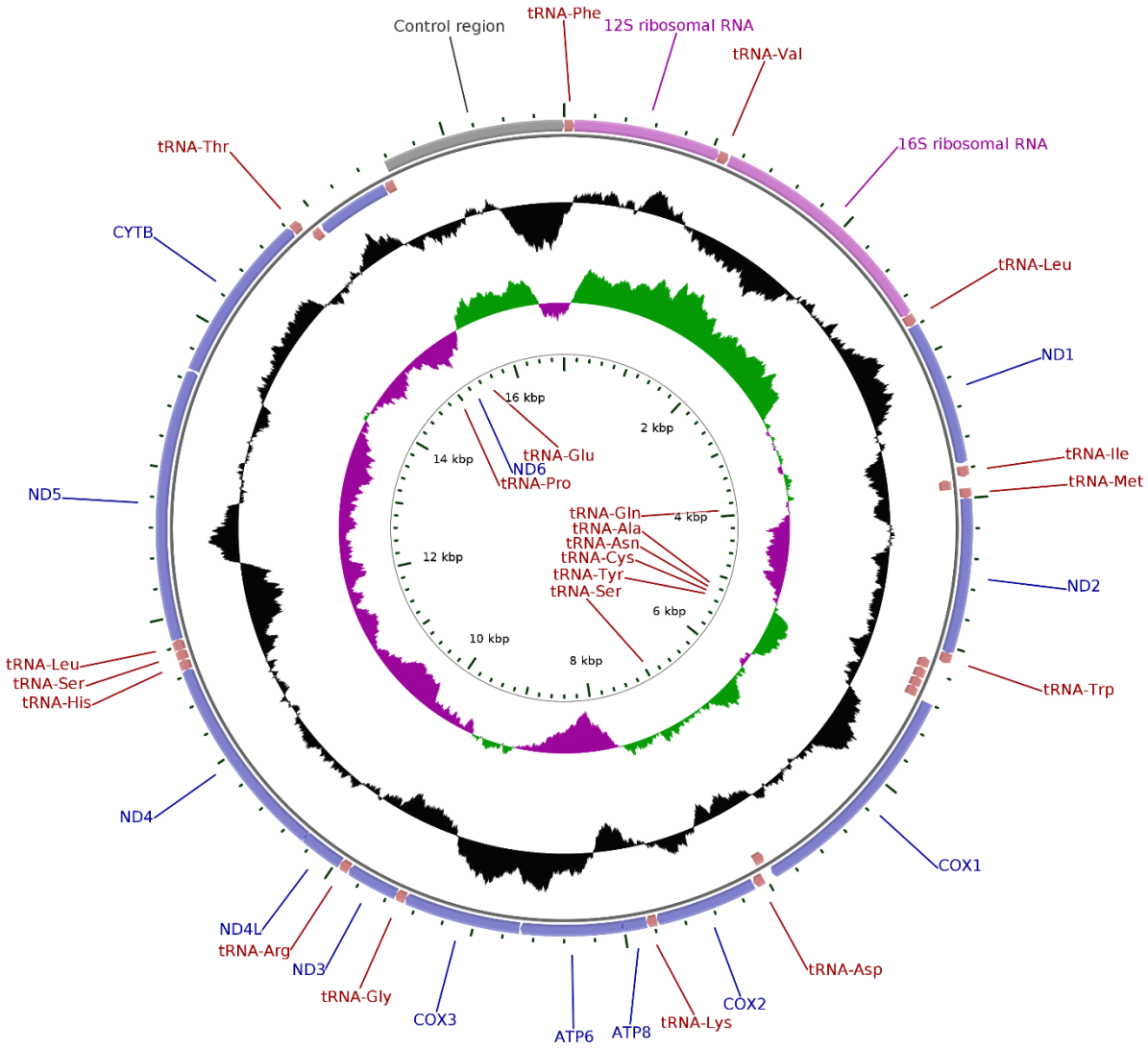

Figure 2
160x145 mm (x DPI)

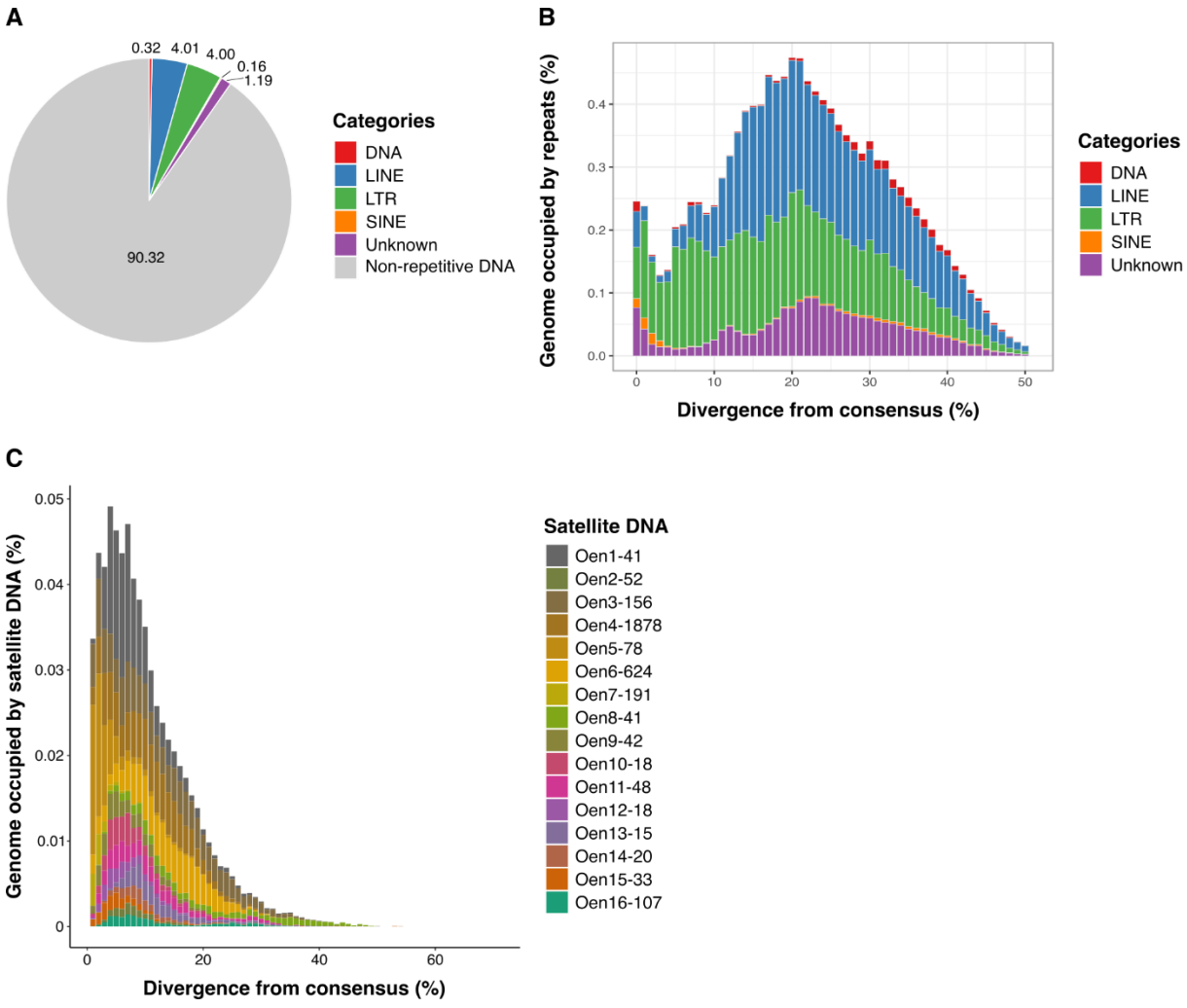

Figure 3
160x135 mm (x DPI)